Audio Encoding of Protein Sequences for Research Utility in Protein Folding Games.

David J. Manglano, University of Chicago

Abstract.

Human-machine interfaces provide important insight into biological systems and are an often-overlooked area of bioinformatic research. Visualization environments provide insight into protein structure and function from sequences rendered as CGI wireframes. Haptic interfaces have been implemented but failed to gain much traction. Emerging virtual reality technologies suggest the potential for bioinformatic interfaces which would allow a researcher to experience simulations of say, protein-protein interactions, as an immersive multi-sensory experience. Generative music seeded by nucleotide and amino sequences is an emerging area of interest in translating bioinformatic data into sensory input. I present Biogroove as a proof-of-concept, a piece of generative music that translates motif data from sequence files to audio files, and suggest ways in which the software might be improved or altered, to have utility in a research context.

Introduction.

Most areas of the physical sciences have converged towards computational analysis. As we investigate the natural world and generate experimental evidence of its properties and the behaviour of its systems, the natural response is to model these systems in silicon. In doing so, the researcher may create create experimental environments which could not otherwise exist, or investigate further the properties discovered through rigorous wet laboratory work.

Of the various areas of computational physical sciences, structural modeling continues to see tremendous interest. The reasons are varied, but generally: structural modelling allows the visually-oriented scientist to learn from data faster than they could with equations, and provides fast insight into complex systems that is not necessarily possible in a laboratory environment. Computational physicists model nano-objects to understand the behaviour and synthesis of hydrocarbon scaffolds before attempting to prototype them in materials laboratories [1]. This logic extends to the biological sciences, where computational chemistry models drive the discovery of pharmaceutical therapies, usually through a non-interactive molecular dynamics modeling approach. Experimentally, this has proven fruitful, with discoveries ranging from RIPK1/RIPK3 inhibitors (which inhibit necrosis pathways) [2], to in silico molecular-docking evidence for the unexpected efficacy of HIV inhibitors as cancer therapies[3].

So then, visual and molecular dynamics modeling have experimental value in both nanomaterials engineering and pharmaceutical engineering, value which could conceivably extend to the difficult field of rational protein design. Viewed from a cognitive context, a good model of a protein should invite the user into what Mihály Csíkszentmihályi called the “flow state,” in which the challenge of finding insights in the model is balanced by the researcher’s knowledge of the model, holding their interest and stimulating curiosity and focus. How do we create an interactive research environment that enables the researcher to maintain this state for the greatest amount of time? Perhaps more immersive models, ones which transcend the visual cortex and engage the entire brain, are the answer.

Background.

This concept is not entirely new; non-visual methods of protein modeling have been tried and met with limited success. The ProFeel method of haptic interaction with protein models

used of-the-shelf force-feedback controllers to provide data on atomic forces within a protein (attraction and repulsion) to the user's hands [4]. This method informally demonstrated efficacy in "disambiguating data which may be unclear if only presented visually" in the process of helping users to discover protein structural alignments, but also saw limited utility in presenting more than two data types haptically, perhaps due to intrinsic limitations on human perception. So then, if a single sensory input has an upper bound on the amount of data it can disambiguate, and the addition of a second sensory data input increases an individual's ability to discern amongst data, then perhaps further engagement of the senses would refine the discernment further.

Consider Foldit, a game designed to harvest algorithms from users to enhance preexisting protein folding algorithms—a sort of crowdsourced evolutionary algorithm in itself. Foldit only provides one set of inputs to the player, a visual representation of a protein from which the player must remove clashes. With this limited interface, Foldit has been successful in its goal, ultimately generating a model from which the structure of a retroviral protease was derived [4]. Setting aside the ethical implications of crowdsourced scientific labours, the success of Foldit raises the question of what scientific insights a better, more immersive environment may provide. An improved Foldit that haptically, acoustically, and visually engages players could provide faster generation of models. Lacking the time and resources to create such a game myself, I instead focused on the question of how we might accomplish acoustic immersion that provides data about protein sequences to the players, and developed a proof of concept.

Methodology & Approach.

Biological sequences contain many repeated motifs that determine the structure and function of their derived constructs, something they have in common with musical compositions. However, the challenge here is to create an audio file that says something from which a player can intuit information about the structure of the protein being folded or aligned. I focused upon the residues which characterized α-helices and β-sheets (MALEK residues, TYVFIW residues, respectively). The Biogroove algorithm accepts a protein sequence as input, and slides along the sequence at 10-amino intervals, analyzing each interval for %MALEK or %TYVFIW content. The helices and beta sheets have associated generative motifs, which modify subsequent motifs—if the protein contains more than one area with a helix or a beta-sheet, this will be evident in the music produced.

Biogroove attempts to create an audio file where, for two given protein sequences where, when sequences are aligned, similar protein motifs will sound resonant to a human being aligning the sequences. The algorithm also modifies the generated music in areas where the %MALEK or %TYVFIW is above the 50% threshold for preceding sequences, to suggest audibly that the discovered subsequence containing motif-significant amino acids is part of a motif and not incidental.

In order to achieve this goal, I used the pdremix library to read in PDB files, and the pysynth library to achieve audio synthesis in python. The Biogroove algorithm can be found at my GitHub (http://github.com/manglano), along with a few examples of protein sounds.

Discussion.

Biogroove uses the standard elements of functional programming to produce generative music, using musical motifs for the two preeminent structural motifs in proteins. The analogy of protein sequences to music is familiar to many researchers. The pattern of meaning-making

through repeated motifs in both music and biology has been recognized by biology researchers [5].

This code achieved the desired result, in that areas of the protein sequence with a majority of amino acids suggesting either an α-helix or a β-sheet are evident from the generated WAV file. Stretches of the sequence containing motifs in sequence alter the generated music such that this information is immediately evident to the listener. The generated music does suggest the presence of helices in areas where helices are present in human hemoglobin subunit alpha.

There are several improvements I would make, given more time to devote to this project (and a Python synthesizer library with multitrack support). First, I would add a bass track generated from the superfamily motifs, using NCBI Conserved Domains. An implementation with research utility might request the conserved domain information from NCBI, parse the output, and generate a different bass track for each conserved domain, modifying these tracks if they belong to the same superfamily. This is not a simple endeavor, but would perhaps be valuable within the context of a game like Foldit.

Another useful way to modify the output might be to use a scoring matrix to determine whether or not a stretch of amino acids is part of a motif, instead of using a heuristic, as in the current implementation. Then, apply a VST effect like distortion or phaser to encode this data as sound. In this manner, we improve the resolution of the data encoded in the audio.

The desired outcome would be to package the output of this algorithm in such a way as to provide useful data to players of a Foldit-like, which would require aligning segments of audio and actively modifying these segments of audio in a way that suggests their RMSD. In order to give information about the alignment of audio-encoded sequences, one could overlay the tracks and modify the perceived distance of a given track using cognitive maps of the human auditory system [6]. Unlike the last two improvements, this implementation is far beyond my expertise, but suggests exciting potential for audio-encoded data.

Conclusion.

The utility of multisensory data abstractions was demonstrated. A proof-of-concept, the Biogroove algorithm, was provided for audio encoding of protein sequence data which provides meaningful information regarding the sequence structure. The algorithmic model was elaborated upon to describe a methodology for the generative construction of audio-encoded “data music” whose acoustic & psychoacoustic features would more accurately describe the structure and 3D spatial alignment of proteins.

Bibliography.

[1] Interactive physically-based structural modeling of hydrocarbon systems. Bosson, Grudinin, et al. Journal of Computational Physics. 2012. doi:10.1016/j.jcp.2011.12.006
[2] Ensembling and filtering: an effective and rapid in silico multitarget drug-design strategy to identify RIPK1 and RIPK3 inhibitors. Fayaz & Rajanikant. J Mol Model. 2015. doi: 10.1007/s00894-015-2855-2
[3] Could the FDA-approved anti-HIV PR inhibitors be promising anticancer agents? An answer from enhanced docking approach and molecular dynamics analyses. Arodola & Soliman. Drug Des Devel Ther. 2015. doi: 10.2147/DDDT.S87653.
[4] High-resolution structure of a retroviral protease folded as a monomer. Girski, Kazmierczyk, et al. Acta Crystallogr D Biol Crystallogr. 2011. doi: 10.1107/S0907444911035943.

[5] The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition. Immunogenetics August 1986, Volume 24, Issue 2, pp 71-78. Ohno & Ohno.
[6] Introduction to Psychoacoustics. Begault, 2011. http://humansystems.arc.nasa.gov/publications/Begault_2002_Introduction_to_Psychoacoustics.pdf